Asymptotically Optimal Model Estimation for Quantization

Alexey Ozerov and W. Bastiaan Kleijn

***Abstract*—Using high-rate theory approximations we introduce flexible practical quantizers based on possibly non-Gaussian models in both the constrained resolution (CR) and the constrained entropy cases. We derive model estimation criteria optimizing asymptotic (with increasing rate) quantizer performance. We show that in the CR case the optimal criterion is different from the maximum likelihood criterion commonly used for that purpose and introduce a new criterion that we call constrained resolution minimum description length (CR-MDL). We apply these principles to the generalized Gaussian scaled mixture model, which is accurate for many real-world signals. We provide an explanation of the reason why the CR-MDL improves quantization performance in the CR case and show that CR-MDL can compensate for a possible mismatch between model and data distribution. Thus, this criterion is of a great interest for practical applications. Our experiments apply the new quantization method to controllable artificial data and to the commonly used modulated lapped transform representation of audio signals. We show that both the CR-MDL criterion and a non-Gaussian modeling have significant advantages.**

***Index Terms*—Constrained resolution, high-rate theory, model-based quantization, asymptotically optimal model estimation, minimum description length, maximum likelihood.**

I. Introduction

HIGH-RATE (HR) theory approximations, as applied to quantization [1], form a powerful tool allowing to derive analytical asymptotic expressions of quantizer performance. These expressions are usually applied in the following contexts:

- *application 1*: to analyse asymptotic behavior of Lloyd-optimal vector quantizers [2], [3],
- *application 2*: to optimize asymptotic performance of some pre-defined structured quantizers [4], [5],
- *application 3*: to build practical quantizers, given some parametric representation θ (also referred to hereafter as *model*) of the data distribution $p_S(s)$ [6], [7].

In this work we are mainly interested by the third application. It facilitates the design of practical quantizers with the following attractive properties:

- *flexibility*: the quantizers can be built in real time for any value of the rate from the continum of values,
- *low storage requirements*: one does not need to store codebooks, only model parameters need to be stored,
- *low computational load*: the computational complexities of both encoder and decoder are low and independent of the particular rate value.

Such flexible quantizers were recently successfully applied to audio coding [8]–[10], but can be applied for coding of any data, e.g., images or video. Moreover, while HR theory is (asymptotically) valid for high rates, flexible quantizers give in practice satisfactory results for low rates as well [9], [10].

To build such model-based flexible quantizers it is usually implicitly assumed that the model is able to represent the data distribution "perfectly", and the maximum likelihood (ML) criterion is generally used for model estimation [6]–[10]. Thus, except for two works [11], [12] (we discuss the novelty of our proposal, as compared to these works below), the question of model estimation is not very carefully addressed in terms of the best rate-distortion (RD) tradeoff, which is the real objective of quantization.

Assuming that the HR theory assumption holds, we are looking in this paper for model estimation strategies leading to the best RD tradeoff. We consider a k-dimensional random source vector S and assume that its distribution admits a probability density function (pdf) $p_S(s)$. Let source vector S be quantized (e.g., as in [6] or [7]) using a probabilistic model $\theta \in \Theta$ from a family of models Θ, characterized by its pdf $f_S(s|\theta)$. The problem of optimal model estimation consists of choosing a particular $\theta^* \in \Theta$ that leads to the best RD tradeoff.

It is implicitly assumed in the state-of-the-art [6]–[10] that there exist $\theta \in \Theta$ such that $f_S(s|\theta) = p_S(s)$. However, this assumption is almost never verified in practice for the following (possibly redundant) reasons:

- one cannot consider an arbitrary parametric family of distributions, since we do not know yet how to build practical flexible quantizers in the most general case,
- one cannot use an arbitrary model order, since model transmission would cost too much [13], or data overfitting would lead to a decrease of overall quantization performance [14],
- and may be most importantly, the real data distribution often does not fit the model distribution in practice, whatever the parametric family.

In summary, the flexible quantizers of application 3 are usually derived based on theoretical results from application 1. However, while in application 1 it is suitable to consider only

A. Ozerov is with METISS team of INRIA, Rennes Bretagne Atlantique, Campus de Beaulieu, F-35042 Rennes Cedex, France. E-mail: alexey.ozerov@inria.fr.

W. B. Kleijn is with School of Engineering and Computer Science, Victoria University of Wellington, P.O. Box 600, Wellington 6140, New Zealand. E-mail: bastiaan.kleijn@ecs.vuw.ac.nz.

This work was supported in part by the European Union under Grant FP6-2002-IST-C 020023-2 FlexCode.

one data distribution $p_S(s)$[1], it is not suitable for application 3, as explained.

The main goal of this work is to compensate for possible mismatch between the data and model distributions during the model estimation step. This goal can be achieved by the following two options: (i) adjust the model, or (ii) adjust the resulting quantizer density. Here we chose following the first option, since, in our opinion, it is the most promising one. To be more precise, our methodology consists of the following steps:

- assume the model family Θ includes the "right data model", i.e., a $\theta \in \Theta$ such that $f_S(s|\theta) = p_S(s)$ exists, and derive, e.g., as in [3], [15], the quantizer *centroid density function* (see Sec. II-A below) expressed in a parametric form (i.e., via θ) that is optimal in terms of the RD tradeoff,
- keep the parametric form of the obtained quantizers, and derive the so called *operational rate-distortion function (RDF)* [2], for example as in [15], but, in contrast to [15] and in line with [16], [17], assuming $f_S(s|\theta) \neq p_S(s)$, i.e., remove the "right data model" assumption,
- optimize model θ such as to have the best RD tradeoff, i.e., minimize the operational RDF.

In other words, in the last step we approach the philosophy of application 2. Indeed, we just consider a family of the quantizers parameterized by $\theta \in \Theta$, we forget about underlying probabilistic model, and we are simply looking for the θ^* optimizing the operational RDF.

We apply the proposed methodology to both constrained entropy (CE) (variable rate) and constrained resolution (CR) (fixed rate) quantizers, assuming a quite general (possibly non-Gaussian) model distribution. Analyzing operational RDFs for both the CE and CR cases we show that the ML criterion results in optimal performance for the CE case but not for the CR case. For the CE case, the result is consistent with the minimum description length (MDL) principle [18], [19]. We call the new model estimation criterion for CR quantization *CR-MDL*. Our framework is quite general and can be applied to a large range of model distributions. In the experimental part we use generalized Gaussian distributions (GGD) and so-called *generalized Gaussian scaled mixture models (GGSMM)* as source models and apply them to synthetic data (sequences sampled from some GGDs) and real data (modulated lapped transform (MLT) coefficients of speech).

Concerning the two abovementioned existing works, Duni and Rao [11] develop a similar CR-quantization framework in a particular case of GMMs, and [12] is our previous contribution, where we also consider optimal parameter estimation for the Gaussian case. Here we formulate our framework first in the case of any model, and then in a practical case of flexible quantizers derived from possibly non-Gaussian distributions including for example GGD, mixtures of GGDs, etc. Both the formulation of flexible quantizers and the derivation of optimal parameter estimation criteria for the non-Gaussian case[3] are new results. Moreover, in contrast to [11], in our experiments we provide a systematic comparison between CE quantization and CR quantization using both the ML and CR-MDL criteria. Finally, in contrast to [11] and [12], our derivations of asymptotically optimal model estimation criteria are based on theoretical mismatch results in high-resolution quantization theory [16], [17].

In summary, this paper includes the following contributions, as compared to the state of the art:

1) Both the CR [6] and the CE [7] probabilistic model-based quantization schemes are extended to a wider class of non-Gaussian models.
2) As compared to [11], [12], asymptotically optimal model estimation criteria are derived in the general case of any model and for the proposed practical non-Gaussian model-based quantizers using theoretical results from [16], [17].
3) The advantages of both non-Gaussian modeling and optimal estimation criteria are demonstrated for quantization of speech MLT coefficients using GGSMM. To our best knowledge, while Gaussian models have been used for quantization of linearly transformed speech coefficients [8]–[10], [12], such non-Gaussian model-based schemes were not yet studied in this context.

The remainder of this paper is organized as follows. A quite general formulation of a model estimation framework is given in section II. However, in this section we do not consider how to build flexible quantizers for such a general case. Thus, in section III the framework is reformulated for the case of practical flexible quantizers, considering a particular class of parametric model families. Experiments on CE and CR GGD model-based quantization of synthetic and real data using different parameter estimation criteria are presented in Section IV. In Section V the proposed framework is discussed and some conclusions are drawn.

II. General formulation

In order to provide a better understanding of our framework, we use a general formulation. However, such a general formulation is not directly applicable for practical flexible quantizers, and the corresponding reformulation will be given in section III.

To derive operational RDFs and their mismatched versions, i.e., when $f_S(s|\theta) \neq p_S(s)$, we here follow the results by Zador [2], Bucklew [16], Gray *et al.* [3], [15], [17] and [20]. Under some assumptions, asymptotic validity of operational RDFs was shown in [2], and in [16] and [17] for the mismatched cases. Lower and upper bounds of achievable performance were studied in [3]. We here leave aside the study of achievable performance bounds, and, instead, are interested in optimizing the quantizer's asymptotic performance, i.e., the

[1]In fact, in application 1 there is no parametric model at all, and non-parametric (and non-flexible) Lloyd-optimal vector quantizer approaches the HR theory optimal quantizer for $p_S(s)$, as rate goes up.

[2]Given a quantizer specified by its centroid density function and some data specified by its distribution, the operational RDF, as introduced in [15], represents the expected RD relation for the quantizer, as applied to the data.

[3]As it will be explained in details later, in this paper we consider GMM-based quantization as quantization using a single Gaussian with parameters varying in time. This is in fact the case, since for quantization of one source vector only one pre-selected Gaussian component is used [6], [7].

operational RDF. The adverb "asymptotically" in the paper's title reflects this point.

A. Quantizers

We consider again the source vector S with data distribution pdf $p_S(s)$ and model distribution pdf $f_S(s|\theta)$ ($\theta \in \Theta$). We first suppose that a $\theta \in \Theta$ exists such that $f_S(s|\theta) = p_S(s)$. Let s be a particular realization of the source vector S, and $\mathcal{Q}(s)$ be its quantized version. For quantization we consider the mean r-th power distortion measure:

$$d_r(s, \mathcal{Q}(s)) = \frac{1}{k} \left\| s - \mathcal{Q}(s) \right\|_r = \frac{1}{k} \left(\sum_{i=1}^{k} (s_i - \mathcal{Q}(s)_i)^2 \right)^{\frac{r}{2}}. \quad (1)$$

Let $\{\mathcal{Q}_m\}_{m=1}^{+\infty}$ be a sequence of quantizers with a total number of reconstruction points $\{L_m\}_m$ (such that $L_m \to +\infty$ while $m \to +\infty$). Assuming these quantizers are optimal for data with pdf $f_S(s|\theta)$, *point density function* $\Lambda(s|\theta)$ is defined as (see e.g., [3]) a continuous function such that for any "reasonable" subset $\mathcal{S} \subset \mathbb{R}^k$ the ratio between reconstruction points in $\mathcal{S}$ and L_m tends to $\int_{\mathcal{S}} \Lambda(s|\theta) ds$ when $m \to +\infty$. Here we use a so called *centroid density function* $g_{C,m}(s|\theta)$ that relates to the to point density function $\Lambda(s|\theta)$ as $g_{C,m}(s|\theta) = L_m \Lambda(s|\theta)$.

It can be shown [3], [20] that the mean distortion $D_m = \mathbb{E}[d(S, \mathcal{Q}_m(S))]$ can be expressed "asymptotically" as:

$$D_m = \int_{\mathbb{R}^k} f_S(s|\theta) C(r, k, \mathcal{G}_k(s)) g_{C,m}(s|\theta)^{-\frac{r}{k}} ds, \quad (2)$$

where $C(r, k, \mathcal{G}_k(s))$ is the normalized moment of inertia or coefficient of quantization [20], and $\mathcal{G}_k(s)$ indicates the geometry of the cell used for quantization of vector s.

More precisely, equation (2) is valid "asymptotically" in the sense that the right part of (2) divided by L_m tends to D_m/L_m when $m \to +\infty$. For the sake of simplicity, we use in (2) and in other expressions below the equality sing ($=$) instead of the approximation ($\approx$). Moreover, and for the same reason, we drop the index m in all expressions below.

Assuming optimal geometry and that Gersho's conjecture [21] holds, i.e., for optimal geometry the normalized moment of inertia does not vary with the cell index ($\mathcal{G}_{\mathrm{opt},k}(s) = \mathcal{G}_{\mathrm{opt},k}$), we can write:

$$D = C_{r,k} \int_{\mathbb{R}^k} f_S(s|\theta) g_C(s|\theta)^{-\frac{r}{k}} ds, \quad (3)$$

where $C_{r,k} = C(r, k, \mathcal{G}_{\mathrm{opt},k})$.

We would like to derive the optimal centroid density function $g_C(s|\theta)$ under the following two constraints on the rate:

1) *Constrained entropy*, when each source vector can be quantized with any number of bits, and only the first-order entropy of the quantization indices is constrained. It can be shown [3], [20] that under HR theory assumptions this constraint is equivalent to:

$$-\int_{\mathbb{R}^k} f_S(s|\theta) \log_2 \frac{f_S(s|\theta)}{g_C(s|\theta)} ds \le R, \quad (4)$$

with R denoting the average rate (in bits per vector).

2) *Constrained resolution*, when each source vector can be quantized with at most R bits, which in terms of centroid density function is equivalent to:

$$\log_2 \int_{\mathbb{R}^k} g_C(s|\theta) ds \le R, \quad (5)$$

with R denoting the constant rate.

To derive optimal centroid density functions one can minimize mean distortion D expressed by Eq. (3) under the corresponding rate constraint ((4) or (5)) using, e.g., the Lagrange multiplier method (see [3], [20]). In the CE case the optimal centroid density is constant and related to the average rate as follows:

$$\log_2 g_C^{\mathrm{opt,CE}}(s|\theta) = R + \int_{\mathbb{R}^k} f_S(y|\theta) \log_2 f_S(y|\theta) dy, \quad (6)$$

and in the CR case the optimal centroid density can be written as:

$$g_C^{\mathrm{opt,CR}}(s|\theta) = 2^R \frac{f_S(s|\theta)^{\frac{k}{k+r}}}{\int_{\mathbb{R}^k} f_S(y|\theta)^{\frac{k}{k+r}} dy}. \quad (7)$$

B. Operational rate-distortion functions

By substituting Eqs. (6) and (7) into Eq. (3), it follows that in both the CR and CE cases and under HR theory assumptions the (average) rate R (in bits per vector) is related to the (average) distortion D (per dimension) via the following so-called *operational RDF*:

$$R = -\frac{k}{r} \log_2 D + \psi(\theta), \quad (8)$$

where in the CE case the term $\psi(\theta)$ is:

$$\psi_{\mathrm{CE}}(\theta) = \frac{k}{r} \log_2 C_{r,k} - \int_{\mathbb{R}^k} f_S(s|\theta) \log_2 f_S(s|\theta) ds, \quad (9)$$

while in the CR case it is:

$$\psi_{\mathrm{CR}}(\theta) = \frac{k}{r} \log_2 \left[C_{r,k} \left(\int_{\mathbb{R}^k} f_S(s|\theta)^{\frac{k}{k+r}} ds \right)^{\frac{k+r}{k}} \right]. \quad (10)$$

Recall that all the derivations above were done under the assumption $f_S(s|\theta) = p_S(s)$ (see Sec. II-A). However, as discussed in the introduction, in the most practical situations the true data density $p_S(s)$ does not belong to the family of model densities $\{f_S(s|\theta)\}_{\theta \in \Theta}$ and can only be approximated by a member from this family ($p_S(s) \approx f_S(s|\theta)$) with more or less success.

C. Mismatched operational rate-distortion functions

Now we relax the assumption $f_S(s|\theta) = p_S(s)$, but we still consider optimal quantizers derived under this assumption (i.e., a uniform quantizer in the CE case and a quantizer with centroid density $g_C^{\mathrm{opt,CR}}(s|\theta)$ (7) in the CR case). Under these assumptions, we are looking for model parameter estimation criteria, that are optimal in terms of quantization performance. The assumption $f_S(s|\theta) \neq p_S(s)$ leads to the replacement of the first entry of $f_S(\cdot|\theta)$ in Eqs. (3) and (6) by $p_S(\cdot)$. Doing that and performing similar derivations, one can find, under certain conditions (see Theorem 2 of [17] and Theorem 2

of [16] or Appendix A), the following operational RDF (analogous to (8)):

$$R = -\frac{k}{r}\log_2 D + \psi(\theta, S), \tag{11}$$

with

$$\psi_{\mathrm{CE}}(\theta, S) = \frac{k}{r}\log_2 C_{r,k} - \int_{\mathbb{R}^k} p_S(s)\log_2 f_S(s|\theta)ds, \tag{12}$$

$$\psi_{\mathrm{CR}}(\theta, S) = \frac{k}{r}\log_2\left[C_{r,k}\frac{\int_{\mathbb{R}^k} p_S(s) f_S(s|\theta)^{-\frac{r}{k+r}}ds}{\left(\int_{\mathbb{R}^k} f_S(y|\theta)^{\frac{k}{k+r}}dy\right)^{-\frac{r}{k}}}\right]. \tag{13}$$

Such *mismatched operational RDFs* (i.e., when an optimal quantizer is derived for model distribution $f_S(s|\theta)$, but applied to data having a different distribution $p_S(s)$), were already reported by Bucklew [16] for the CR case and by Gray and Linder [17] for the CE case. Moreover, these works provide rigorous mathematical conditions that are sufficient for asymptotic validity of (11), (12) and (13). Here, we use these results for model estimation.

D. Optimal model estimation

We see from equation (11) that under HR theory assumptions the mismatched operational RDF (for both CR and CE cases) is a linear function with slope $-k/2$ and intercept $\psi(\theta, S)$, relating the rate and the logarithm of distortion. Thus, to minimize the distortion D for any (high) rate R, one must look for model parameters θ minimizing the term $\psi(\theta, S)$, which is equivalent to, respectively for the CE and CR case,

$$\theta_{\mathrm{CE}}^{\mathrm{opt}} = \arg\max_{\theta}\int_{\mathbb{R}^k} p_S(s)\log f_S(s|\theta)ds, \tag{14}$$

$$\theta_{\mathrm{CR}}^{\mathrm{opt}} = \arg\min_{\theta}\frac{\int_{\mathbb{R}^k} p_S(s) f_S(s|\theta)^{-\frac{r}{k+r}}ds}{\left(\int_{\mathbb{R}^k} f_S(y|\theta)^{\frac{k}{k+r}}dy\right)^{-\frac{r}{k}}}. \tag{15}$$

Note that the criteria for estimation of the optimal model distribution in the CE and the CR cases are different.

E. Case of empirical data distribution

In many practical situations we do not know the true data distribution (i.e., $p_S(s)$), and we have only a sequence of observed vectors $\mathbf{s} = \{s^n\}_{n=1}^N$ ($s^n \in \mathbb{R}^k$) that we would like to quantize. In that case one can obtain the following *empirical mismatched RDF* (see Appendix A for derivations):

$$R = -\frac{k}{r}\log_2 D + \psi^{\mathrm{emp}}(\theta, \mathbf{s}), \tag{16}$$

with

$$\psi_{\mathrm{CE}}^{\mathrm{emp}}(\theta, \mathbf{s}) = \frac{k}{r}\log_2 C_{r,k} - \frac{1}{N}\log_2\prod_{n=1}^{N} f_S(s^n|\theta), \tag{17}$$

$$\psi_{\mathrm{CR}}^{\mathrm{emp}}(\theta, \mathbf{s}) = \frac{k}{r}\log_2\left[C_{r,k}\frac{\frac{1}{N}\sum_{n=1}^{N} f_S(s^n|\theta)^{-\frac{r}{k+r}}}{\left(\int_{\mathbb{R}^k} f_S(y|\theta)^{\frac{k}{k+r}}dy\right)^{-\frac{r}{k}}}\right]. \tag{18}$$

In contrast to (11), which requires knowledge of the underlying probability distribution, the operational rate distortion relation (16) is useful for real-world data. It predicts the rate-distortion relation for a set of N data points $\mathbf{s} = \{s^n\}_{n=1}^N$ for the case that the signal model $f_S(\cdot|\theta)$ is assumed.

The optimal model estimation criteria (analogous to (14) and (15)) become:

$$\theta_{\mathrm{CE}}^{\mathrm{opt}} = \theta_{\mathrm{ML}} = \arg\max_{\theta}\prod_{n=1}^{N} f_S(s^n|\theta), \tag{19}$$

$$\theta_{\mathrm{CR}}^{\mathrm{opt}} = \theta_{\mathrm{CR\_MDL}} = \arg\min_{\theta}\frac{\sum_{n=1}^{N} f_S(s^n|\theta)^{-\frac{r}{k+r}}}{\left(\int_{\mathbb{R}^k} f_S(y|\theta)^{\frac{k}{k+r}}dy\right)^{-\frac{r}{k}}}. \tag{20}$$

Thus, in the CE case the ML criterion is optimal in terms of quantization performance, which is consistent with the minimum description length (MDL) principle [18], [19]. However, in the CR case we have an optimal model estimation criterion that in general is not equivalent to ML. We call this new criterion *CR-MDL*.

F. Discussion

Unfortunately, except in the scalar case ($k = 1$), we do not know how to design analytically practical flexible coders (including the quantization and the indexing) in the above-described general situation[4]. As a result, the Gaussian model is usually considered in practice (see, e.g., [9], [10]), i.e., $f_S(s|\theta)$ is set to be Gaussian. The more general GMMs are considered in [6]–[8], [22]. However, GMM-based quantization consists of selecting a suitable Gaussian component and using only this component for quantization, which results in loss of optimality when the components overlap [7]. In other words this quantization is locally Gaussian. Thus, while we are aware that GMMs can approach any distribution with more or less success, we consider here GMM-based quantization as quantization using a Gaussian model with time varying parameters.

The most common approach to build flexible coders in this case is to first decorrelate quantized source vector using the Karhunen-Loeve transform (KLT), and then quantize the vector components independently using corresponding scalar quantizers (see [7] for CE case and [6] for CR cases). For such schemes, the memory advantage of vector quantization versus scalar quantization (see, e.g., [20], [23]) is taken into account because of the KLT. However, the space filling advantage and the shape advantage (for the CR case) are not used. For the CE case Zhao *et al.* [7] proposed also using general lattices instead of Z-lattices (scalar quantizers) in the KLT domain, and the resulting scheme takes into account the space filling advantage. The situation is more complex in the CR case, one approach taken was to apply scalar companders and general lattices in the KLT domain [22] (instead of scalar quantization [6]), but the centroid density of such a quantizer can be far from the optimal centroid density (7), which in principle cannot be implemented via scalar companders [24].

[4]More precisely, in such a general situation, the quantization is difficult, but not the indexing, for the CR case, and the indexing is difficult, but not the quantization, for the CE case.

For the sake of simplicity and consistency between the CR and CE cases we here consider scalar quantizers in some transformed domain (e.g., as in [7] and [6]), but extend them to a more general case of non-Gaussian distributions.

III. Practical flexible quantizers

We consider an N-length sequence $\mathbf{S} = \{S^n\}_{n=1}^N$ of k-dimensional real-valued source vectors[5], and the corresponding sequence of observations $\mathbf{s} = \{s^n\}_{n=1}^N$ ($s^n \in \mathbb{R}^k$) to quantize.

A. Source model

Let a source vector S^n be modeled by a distribution with pdf:

$$f_{S^n}(s|\theta_n) = \prod_{i=1}^{k} \lambda_{n,i}^{-1/2} \eta\left([\Lambda_n^{-1/2} U_n^T (s-\mu_n)]_i\right), \quad (21)$$

where μ_n is a vector, U_n is an orthogonal matrix ($U_n^T U_n = I$), $\Lambda_n = \mathrm{diag}\{\lambda_{n,1}, \ldots, \lambda_{n,k}\}$ is a diagonal matrix, and $\eta(\cdot)$ is a scalar pdf. In other words, we assume that after some translation (by μ_n), rotation (by U_n^T) and dimension-wise scaling (by $\Lambda_n^{-1/2}$), the components X_i^n ($i = 1, \ldots, k$) of the resulting random vector $X^n = \Lambda_n^{-1/2} U_n^T (S^n - \mu_n)$ are independent and identically distributed (i.i.d.) with pdf:

$$f_{X_i^n}(x_i|\theta_n) = \eta(x_i). \quad (22)$$

Note that the samples S_i^n are generally not distributed with pdf $\eta(\cdot)$ (up to some scaling and shift). The Gaussian case forms an exception on this rule. Given the pdf $\eta(\cdot)$, such a source model can be parameterized as:

$$\theta \triangleq \{\theta_n\}_{n=1}^N \triangleq \{\mu_n, U_n, \Lambda_n\}_{n=1}^N. \quad (23)$$

Let us remark that estimation of such a model, assuming all parameters are free, is not efficient, since there are more parameters than data samples and such an estimation would lead to a serious data overfitting. Thus, there should be some additional structure that reduces the number of free parameters. For example, one can assume that the set of model parameters is limited to $\{\tilde{\theta}_q\}_{q=1}^Q$ ($Q << N$) and that they are shared between several observations, i.e., $\theta_n = \tilde{\theta}_{q(n)}$ (e.g., as for GMMs [6], [7]). In that case the source vectors $\{S^n | q(n) = q\}$ are i.i.d. and the estimation becomes reliable if the set $\{n|q(n) = q\}$ is sufficiently large. Particular model structures will be specified in the experimental section IV, and we do not do so at that level of presentation for the sake of generality.

B. Practical quantization schemes

For quantization we consider the average mean squared-error (MSE) (a particular case of r-th power distortion measure (1) with $r = 2$):

$$d_2(s, \mathcal{Q}(s)) = (1/k) \left\| s - \mathcal{Q}(s) \right\|_2, \quad (24)$$

which is a *single letter* distortion measure, i.e., for a vector it equals to the mean of the distortions for the vector components. We consider a quantization scheme based on scalar quantization of the independent components that can be summarized as follows:

1) Transform vector s^n into the "independent" domain:
$$y^n = U_n^T (s^n - \mu_n). \quad (25)$$
2) Quantize each dimension y_i^n with a scalar quantizer:
$$Q_{\Lambda_n, \eta(\cdot)}^{Y_i} : y_i^n \to \hat{y}_i^n, \quad (26)$$
that is optimal for the i-th dimension of source $Y^n = U_n^T (S^n - \mu_n)$ under one of the rate constraints (CR or CE), assuming that the HR theory assumptions are valid.[6]
3) Transmit codeword index of $\hat{y}_i^n$ to the decoder together with side information about model parameters $\theta_n = \{\mu_n, U_n, \Lambda_n\}$, that can be quantized as well (if necessary).
4) Reconstruct the quantized vector: $\hat{s}^n = U_n \hat{y}^n + \mu_n$.

The presented quantization scheme is a generalization of several model-based quantization schemes, such as GMM-based quantization [6], [7] (we consider GMM-based quantization as Gaussian model-based quantization, see Sec. II-F), autoregressive model-based quantization [9], [10], and GGD-based flexible quantization that we would like to explore in the experimental part of this paper. Note that the GGD model was already used for quantization (e.g., in [25]). However, the quantizers used in [25] are not flexible, since they are based on Lloyd-Max scalar quantization.

C. Optimal scalar quantizers

In this section we derive expressions for optimal (in terms of minimal overall MSE) scalar quantizers $Q_{\Lambda_n, \eta(\cdot)}^{Y_i}$ (26) for both the CE and CR cases.

1) Constrained entropy: For the CE case with MSE distortion, uniform quantization is asymptotically optimal [1]. Thus, $Q_{\Lambda_n, f(\cdot)}^{Y_i}$ is a scalar quantizer with a constant step size Δ. Using an arithmetic coder as an entropy coder of the codeword indices, the effective codeword length $\mathcal{L}_n$ (in bits) is:

$$\mathcal{L}_n = -\sum\nolimits_{i=1}^{k} \log_2 \int_{\hat{y}_i - \Delta/2}^{\hat{y}_i + \Delta/2} f_{Y_i^n}(y_i) dy_i, \quad (27)$$

where

$$f_{Y_i^n}(y_i) = \lambda_{n,i}^{-1/2} \eta(y_i \lambda_{n,i}^{-1/2}) \quad (28)$$

is the model pdf of the i-th component of vector $Y^n = U_n^T (S^n - \mu_n)$.

[5] In contrast to the previous section we assume here that the random vector is dependent on the index n. This is because we want the model (as will be introduced below) be dependent on n.

[6] Given that X_i^n are i.i.d. with pdf $\eta(\cdot)$ (22) and $Y_i^n = \lambda_{n,i}^{1/2} X_i^n$, the resulting expressions for the optimal scalar quantizers $Q_{\Lambda_n, f(\cdot)}^{Y_i}$ are indeed independent of μ_n and U_n, since the MSE distortion measure (24) is invariant under the transform $U_n^T(\cdot - \mu_n)$, as a result of the orthogonality of U_n.

2) Constrained resolution: Let $R_{n,i}$ be the number of bits spent for i-th dimension of the n-th vector. Since the MSE distortion (24) is a single letter distortion the scalar quantizer $Q^{Y_i}_{\Lambda_n,\eta(\cdot)}$ must minimize the MSE of the i-th dimension. According to (7) (for $k=1$) such an optimal scalar quantizer (under HR assumptions) has the following centroid density:

$$g_{n,i}(y_i) = L_{n,i}\frac{f_{Y_i^n}(y_i)^{\frac{1}{3}}}{\int_{\mathbb{R}} f_{Y_i^n}(z_i)^{\frac{1}{3}}dz_i}, \tag{29}$$

where $L_{n,i}=2^{R_{n,i}}$ is the number of levels, and $f_{Y_i^n}(y_i)$ is given by (28). Substituting (29) into (3) (for $k=1$) one can write the average MSE distortion for the i-th component of the n-th vector:

$$D_{n,i} = \frac{C_s}{L_{n,i}^2}\left(\int_{\mathbb{R}} f_{Y_i^n}(z_i)^{\frac{1}{3}}dz_i\right)^3 \tag{30}$$

where $C_s = C_{2,1} = 1/12$ is the coefficient of quantization of a scalar quantizer Since the MSE distortion is single letter, the average MSE distortion for the vector Y^n is $D_n = \frac{1}{k}\sum_{i=1}^k D_{n,i}$.

In order to find $L_{n,i}=2^{R_{n,i}}$ we minimize MSE distortion D_n under the rate constraint $\sum_{i=1}^k R_{n,i} \le R$. By using the Lagrange multiplier method we find:

$$\log_2 L_{n,i} = R_{n,i} = \frac{1}{2}\log_2 I_{n,i} + \frac{1}{k}\left[R - \sum_{l=1}^k \frac{1}{2}\log_2 I_{n,l}\right], \tag{31}$$

with $I_{n,i} = \left(\int_{\mathbb{R}} f_{Y_i^n}(z_i)^{\frac{1}{3}}dz_i\right)^3$. Using (28), equations (29) and (31) can be rewritten as:

$$g_{n,i}(y_i) = L_{n,i}\frac{\eta(y_i\lambda_{n,i}^{-1/2})^{\frac{1}{3}}}{\int_{\mathbb{R}}\eta(z_i\lambda_{n,i}^{-1/2})^{\frac{1}{3}}dz_i}, \tag{32}$$

$$\log_2 L_{n,i} = R_{n,i} = \frac{R}{k} + \frac{1}{2}\log_2\left(\lambda_{n,i}/\prod_{l=1}^k \lambda_{n,l}^{1/k}\right). \tag{33}$$

We see that equation (33) is identical to that arrived in [6][7] for a Gaussian pdf $\eta(\cdot)$. So, this expression is valid for any scalar pdf $\eta(\cdot)$; it is independent of the particular form of $\eta(\cdot)$, and depends only on Λ_n and total rate R. In other words, that means that for a single letter distortion measure the bit allocation between scalar CR quantizers having up to some scaling the same point density would be independent of the particular form of this density.

Finally, the scalar quantizer $Q^{Y_i}_{\Lambda_n,\eta(\cdot)}$ with centroid density (32) can be implemented via companding[8] as follows:

1) Compute $x_i = y_i/\sqrt{\lambda_{n,i}}$.
2) Apply the optimal scalar compressor corresponding to the pdf $\eta_{\frac{1}{3}}(\cdot)$ ($\eta_{\frac{1}{3}}(x_i) \triangleq \eta(x_i)^{\frac{1}{3}}/\int_{\mathbb{R}}\eta(z_i)^{\frac{1}{3}}dz_i$):
$$u_i = \xi_{\frac{1}{3}}(x_i),$$
where $\xi_{\frac{1}{3}}(\cdot)$ is the cumulative distribution function (cdf) of a random variable with pdf $\eta_{\frac{1}{3}}(\cdot)$ (i.e., $\xi_{\frac{1}{3}}(x_i) = \int_{-\infty}^{x_i}\eta_{\frac{1}{3}}(z_i)dz_i$).
3) Quantize u_i with a scalar quantizer $Q^{U_i}_{L_{n,i}} : u_i \to \hat{u}_i$ uniform on the interval $(0,1)$ with $L_{n,i}$ levels computed using (33).
4) Reconstruct $\hat{y}_i = \sqrt{\lambda_{n,i}}\xi^{-1}_{\frac{1}{3}}(\hat{u}_i)$.

[7]Note that our derivations are almost the same as in [6], with difference that we do not assume that orthogonal transform U_n^T is the KLT and that $\eta(\cdot)$ is a Gaussian pdf.

[8]Note that companding is optimal for the scalar case.

D. Mismatched operational rate-distortion functions

We consider a sequence of vectors $\mathbf{s} = \{s^n\}_{n=1}^N$, and we assume that these vectors are quantized under HR theory assumptions as described in sections III-B and III-C using a model $\theta = \{\mu_n, U_n, \Lambda_n\}_{n=1}^N$. One can show that in this case the mismatched operational RDF (analogous to (11)) can be written as:

$$R = -\frac{k}{2}\log_2 D + \psi^{\text{flex}}(\theta, \mathbf{s}), \tag{34}$$

with

$$\psi^{\text{flex}}_{\text{CE}}(\theta,\mathbf{s}) = \frac{k}{2}\log_2 C_s - \frac{1}{N}\log_2\prod_n f_{S^n}(s^n|\theta), \tag{35}$$

$$\psi^{\text{flex}}_{\text{CR}}(\theta,\mathbf{s}) = \frac{k}{2}\log_2\left[C_s\left(\int_{\mathbb{R}}\eta(z_i)^{\frac{1}{3}}dz_i\right)^2\right] + \\ + \frac{k}{2}\log_2\frac{1}{kN}\sum_{n=1}^N|\Lambda_n|^{\frac{1}{k}}\sum_{i=1}^k\eta\left(y_i^n/\sqrt{\lambda_{n,i}}\right)^{-\frac{2}{3}}, \tag{36}$$

where $y^n = U_n^T(s^n - \mu_n)$. For the CE case (Eqs (34), (35)) this result is a straightforward consequence of (27). A derivation of the result for the CR case (Eqs (34), (36)) is given in Appendix B.

E. Optimal model parameter estimation

As before, we see from equations (34) and (35) that in the CE case and under HR theory assumptions the ML criterion is optimal in terms of quantization performance, and that this is not true in the CR case. Thus, in the case of flexible CR quantization we introduce the following new model estimation criterion:

$$\theta^{\text{flex}}_{\text{CR\_MDL}} = \arg\min_\theta \phi(\theta,\mathbf{s}), \tag{37}$$

where the term $\phi(\theta,\mathbf{s})$ defined as

$$\phi(\theta,\mathbf{s}) = \sum_{n=1}^N|\Lambda_n|^{\frac{1}{k}}\sum_{i=1}^k\eta\left(y_i^n/\sqrt{\lambda_{n,i}}\right)^{-\frac{2}{3}}, \tag{38}$$

is obtained from the term $\psi^{\text{flex}}_{\text{CR}}(\theta,\mathbf{s})$ (36) by some simplifications such that the new criterion (37) is equivalent to minimizing the term $\psi^{\text{flex}}_{\text{CR}}(\theta,\mathbf{s})$.

IV. Experiments

The goals of the experiments presented in this section are: (a) to check whether the rate and distortion of the practical flexible quantizers follow the theoretically predicted asymptotic behaviour at high rates, (b) to see in the CR case and for different situations, which improvement can be obtained using the optimal CR-MDL criterion, as compared to the ML criterion (as in [6]), for high and low rates, (c) to

investigate, whether the newly proposed non-Gaussian model-based quantizers (with parameters optimized via asymptotically optimal criteria) applied to some real data can bring an improvement, as compared to the Gaussian model-based quantizers [7], [11]. For that, we first provide some results on quantization of synthetic sources, i.e., when we know exactly the distribution the data were sampled from. Then, we provide some practically useful results on quantization of MLT coefficients of speech.

As non-Gaussian source models we use either centered GGDs or their mixtures. The pdf of a centered GGD with shape parameter ν and standard deviation σ can be written as:

$$f_{\mathrm{GGD}}(s|\nu,\sigma) = \frac{\nu\alpha(\nu)}{2\sigma\Gamma(1/\nu)} \exp\left[-\left|\alpha(\nu)\frac{s}{\sigma}\right|^{\nu}\right], \qquad (39)$$

where $\alpha(\nu) = \left[\frac{\Gamma(3/\nu)}{\Gamma(1/\nu)}\right]^{1/2}$, and $\Gamma(\cdot)$ denotes the *Gamma function* defined as: $\Gamma(z) = \int_0^{+\infty} t^{-1+z}e^{-t}\,dt$.

A. Synthetic scalar sources

In the simple case of scalar quantization ($k = 1$) we assume that

- the data sequence follows a GGD with unit variance and the shape factor ν_{data}:

$$p_S(s) = f_{\mathrm{GGD}}(s|\nu_{\mathrm{data}}, 1), \qquad (40)$$

- the model parametric family of pdfs consists of GGD pdfs with the same shape factor ν_{model} ($\nu_{\mathrm{model}} \neq \nu_{\mathrm{data}}$ in general) and different standard deviations (i.e., $\theta = \{\sigma\}_\sigma$):

$$\{f_S(s|\theta)\}_\theta = \{f_{\mathrm{GGD}}(s|\nu_{\mathrm{model}}, \sigma)\}_\sigma, \qquad (41)$$

and we simulate the following three synthetic examples with different degrees of mismatch between the true data distribution and the family of model distributions:

- *Example 1:* $\nu_{\mathrm{data}} = 2$ (Gauss.), $\nu_{\mathrm{model}} = 1$ (Laplacian),
- *Example 2:* $\nu_{\mathrm{data}} = \nu_{\mathrm{model}} = 1.5$ (no mismatch),
- *Example 3:* $\nu_{\mathrm{data}} = 1$ (Laplacian), $\nu_{\mathrm{model}} = 2$ (Gauss.).

1) Implementation issues: In the simulations presented below we optimized the ML criterion (19) and the CR-MDL criterion (37) with respect to (w.r.t.) to the parameter σ. Since, in contrast to the ML criterion, the CR-MDL criterion has no closed-form solution for this model, we used either Newton's method or a gradient descent algorithm, depending on the criterion convexity (in the case of the GGDs the criterion is not always convex). Some implementation details about the quantization and the CR-MDL criterion optimization are given in appendices C-A and C-B, respectively.

2) Simulations: For each of three examples considered the following was performed. A data sequence $\mathbf{s} = \{s^n\}_{n=1}^N$ of length $N = 1000000$ was drawn from pdf $p_S(s)$. Model parameters, denoted as θ_{ML} and $\theta^{\mathrm{flex}}_{\mathrm{CR\_MDL}}$, were estimated using criteria (19) and (37) respectively. Data histograms and estimated model pdfs are represented on the top row of Fig. 1. The data sequence $\mathbf{s}$ was quantized for different rates between 0 and 30 bps in the following three scenarios:

(i) **CR-ML**: CR quantization using model θ_{ML} estimated with the ML criterion,
(ii) **CR-OPT**: CR quantization using model $\theta^{\mathrm{flex}}_{\mathrm{CR\_MDL}}$ estimated with the CR-MDL criterion,
(iii) **CE-OPT**: CE quantization using model θ_{ML} estimated with the ML criterion.

The bottom row of Fig. 1 show the experimental and theoretically predicted (via Eq. (34)) results relatively to the CE-OPT theoretical performance.

3) Discussion: One can note from Fig. 1 that the experimental results follow the theoretically predicted asymptotic behaviour starting from some high rate (20 bps). Performance improvement obtained using optimal CR-MDL criterion, as compared to ML, is huge for the third example (about 40 dB in distortion), moderate, but still important, for the first example, and, as expected, there is no improvement for the second example. In fact, when there is no mismatch between data and model distributions, both criteria should lead to the same parameter estimation. Note also that for the third example the asymptotic behaviour of CR quantization with ML-estimated model is very poor, even if the rate is high (20 bps). This is probably because a heavy-tailed data distribution is modeled by an ML-estimated light-tailed distribution leading to very large quantization cells, i.e., the HR assumptions are violated. The CR-MDL criterion makes the asymptotic behaviour of CR quantization significantly better. Finally, we note that the CR-MDL criterion brings as well some improvement, as compared to ML, for low rates (e.g., starting from 5 bps).

B. Real multidimensional sources: Quantising speech MLT coefficients with GGSMM

In this section we investigate the CR-MDL criterion in the case of real (non-synthetic) multidimensional sources, i.e., when we do not know the "real" data distribution. We consider quantization of modulated lapped transform (MLT) coefficients of speech. Gaussian models are usually used to encode discrete Fourier transform (DFT) [8], MLT [12] or time-domain [9], [10] coefficients of speech. Here we would like to check whether using non-Gaussian (e.g., Laplacian) models for quantization of MLT coefficients of speech can be more advantageous, as compared to Gaussian models. Our motivation is based on our preliminary study [12] and on some works on speech enhancement [26] and separation [27] showing that using Laplacian distributions for speech DFT coefficients can be more advantageous, as compared to Gaussian distributions. More precisely, we consider a so-called *generalised Gaussian scaled mixture model (GGSMM)*. To our best knowledge, such non-Gaussian models were not yet studied in application to quantization of linearly transformed speech samples.

1) GGSMM and coding scheme: Let $\mathbf{s} = \{s^n\}_{n=1}^N$ be a sequence k-dimensional MLT vectors to be quantized. Each vector is assumed to be a realization of a source vector S_n with pdf:

$$f_{S_n}(s|\theta_n^{\mathrm{est}}, \theta^{\mathrm{fix}}) = \prod_{i=1}^{k} f_{\mathrm{GGD}}(s|\nu, h_n\sigma_{q(n),i}), \qquad (42)$$

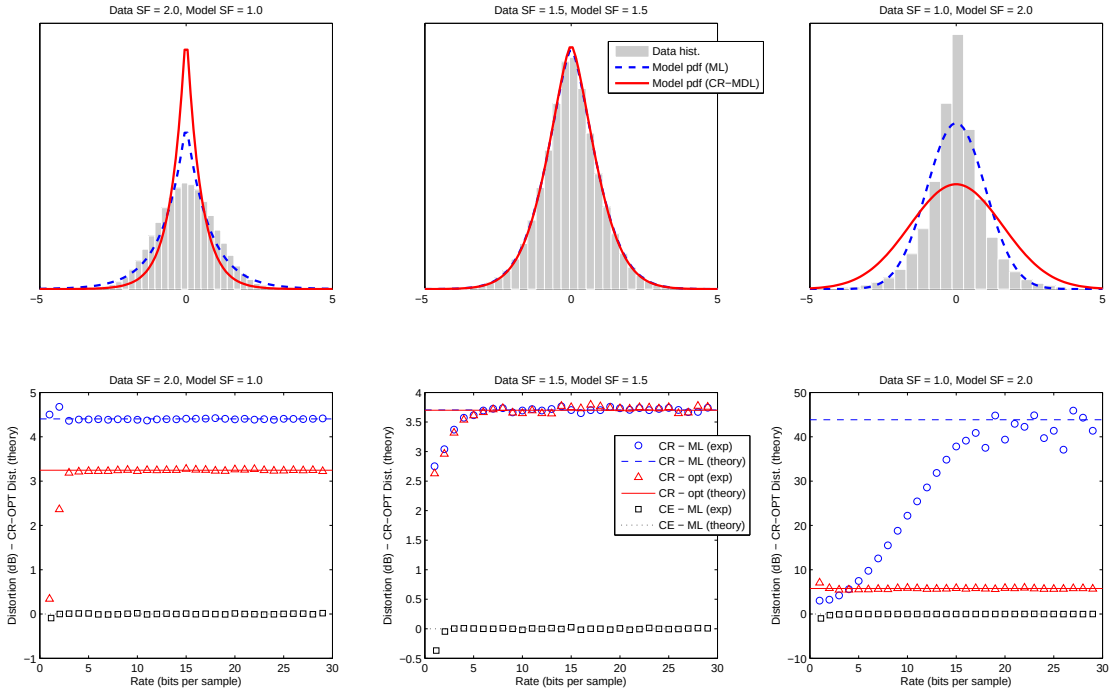

Fig. 1. Results on data sampled from GGDs with shape factors (SFs) $\nu_{\text{data}} = 2, 1.5, 1$ and quantized by GGD-based quantizers with SFs $\nu_{\text{model}} = 1, 1.5, 2$. **Top row**: data histograms (gray bars), ML-estimated model pdfs (blue dashed lines), CR-MDL-estimated model pdfs (red solid lines). **Bottom row**: Experimental results for a set of rates between 0 and 30 bps (circles, triangles or squares) and HR theory predicted RD curves given by equation (34) (lines). The following three scenarios were considered: (i) **CR-ML** (circles and dashed line), (ii) **CR-OPT** (triangles and solid line), (iii) **CE-OPT** (squares and dotted line). All the results are plotted relative to the CE-OPT theoretical performance.

where $\sigma_q = [\sigma_{q,i}]_{i=1}^k$ $(q = 1, \ldots, Q)$ are so-called *characteristic spectral patterns*, $q(n)$ is the index of a spectral pattern selected for n-th MLT vector, and h_n is a non-negative gain accounting for vector's energy. In terms of notations of equation (21) we have $U_n = \mathbb{I}_k$ ($\mathbb{I}_k$ being the $(k \times k)$ identity matrix), $\mu_n = 0$, $\lambda_{i,n} = h_n^2 \sigma_{q(n)}^2$, and $\eta(\cdot) = f_{\text{GGD}}(\cdot|\nu, 1)$. As for coding scheme, parameters $\theta_n^{\text{est}} \triangleq \{q(n), h_n\}$ are estimated for every vector, quantized if necessary, and transmitted to the decoder, and a so-called *dictionary of characteristic spectral patterns* $\theta^{\text{fix}} \triangleq \{\sigma_q\}_{q=1}^Q$ is fixed and supposed to be known, once estimated in a training phase, by both the coder and the decoder. To allow reconstruction of the encoded MLT vector at the decoder, component index $q(n)$ and gain h_n need to be transmitted as well. The index $q(n)$ is losslessly encoded, and the logarithm of gain h_n is lossy encoded using a single Gaussian model and the same HR quantization strategy. As we have found in [13], the asymptotically optimal rate for gain (or more generally model) quantization is fixed, i.e., it is independent on the overall rate.

2) Data and parameters: For evaluation and training, we used respectively 100 and 360 narrow-band speech signals (5 and 15 minutes of speech) randomly selected from respectively the evaluation and training sets of the TIMIT database. The MLT was computed with offset $k = 128$ (16 ms). Finally, we had about 20000 and 60000 MLT vectors for evaluation and training, respectively.

3) Parameter optimization: For GGSMM-based coding scheme we are interested in comparing coding scenarios (i) - (iii) described in section IV-A for different values of the shape factor ν and of the number of model components Q. In order to provide a fair comparison, all the parameters without exception are re-trained for every particular configuration defined by the triple $((l), \nu, Q)$. To optimize parameters for training (θ^{fix} and $\theta^{\text{est}} \triangleq \{\theta_n^{\text{est}}\}_n$) or coding ($\theta^{\text{est}}$ only), we used an iterative procedure consisting in updating in turn a subset of parameters (gains, characteristic spectral patterns, or component indices), given other parameters fixed.[9] As for updates used for gains $\{h_n^{q(n)}\}_{n,q(n)}$ and characteristic spectral patterns $\{\sigma_q\}_q$, the corresponding optimization sub-problems for the ML criterion allow closed form solutions, and for the CR-MDL criterion we used one iteration of Newton's method or gradient descent algorithm, as in section IV-A. The first and second derivatives of the corresponding criterion are quite similar in spirit to those presented in appendix C-B for the scalar GGD case, and they are omitted here for brevity.

Note that the studied coding scheme in the CR case is

[9]Such an optimization procedure is more in line with the segmental K-means algorithm [28] for GMMs rather than with the Expectation-Maximization (EM) algorithm [29] (as e.g., used in [6]–[8]). In our opinion such a way of model training (i.e., when we look for the optimal sequence of component indices, instead of integrating over all possible sequences, as in EM) is more consistent with the common coding strategy [6], [7], where every vector is quantized using only one mixture component.

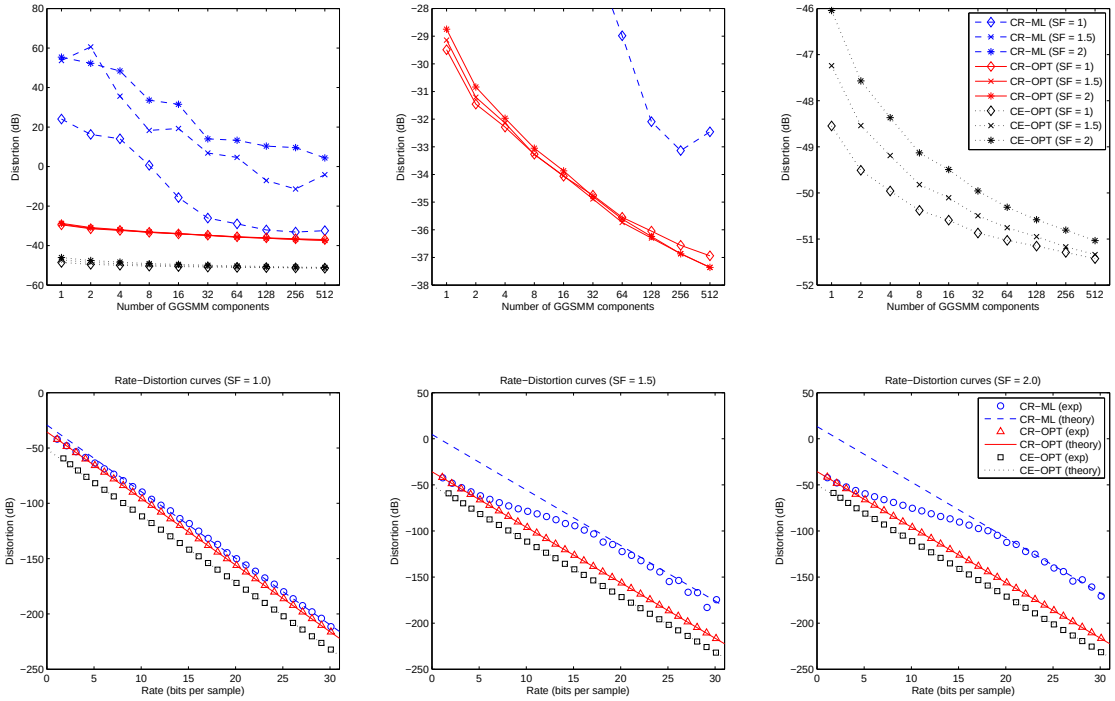

Fig. 2. **Top row:** Zero-rate distortion for different numbers of GGSMM components, scenarios and shape factors (left), zoom on the CR curves (middle), zoom on the CE curves (right). Scenarios: (i) CR-ML (dashed line), (ii) CR-OPT (solid line), and (iii) CE-OPT (dotted line). Shape factors (SF): $\nu = 1$ (diamonds), $\nu = 1.5$ (x-marks), and $\nu = 2$ (stars). **Bottom row:** Experimental results for a set of rates between 1 and 30 bps (circles, triangles or squares) and HR theory predicted RD curves given by equation (34) (lines). Scenarios: (i) CR-ML: (circles and dashed line), (ii) CR-OPT (triangles and solid line), (iii) CE-OPT (squares and dotted line). Shape factors: $\nu = 1$ (left), $\nu = 1.5$ (middle), and $\nu = 2$ (right).

not entirely consistent with conventional GMM-based CR quantization, as described in, e.g., [6]. In fact, in [6] every source vector is quantized with every component, and the component leading to the lowest distortion is selected then, while we are using either the ML or the CR-MDL criterion for component selection. The former strategy is obviously the optimal one, but it also means that the selected model parameter (we consider that the sequence of component indices forms a part of the model) depends on the rate and the quantizer implementation. Since we here prefer staying in the rate-independent model estimation scenario, we leave aside this "optimal component selection strategy" for component selection, and continue using the ML or the CR-MDL criterion. However, we performed some experiments using this "optimal component selection strategy", and noticed that it does not improve the results drastically and does not alter our conclusions on the comparison between ML and CR-MDL criteria. For example, in the case of GSMM with 64 components (the case we study below, that is represented on the bottom right subfigure of figure 2) the "optimal component selection strategy" combined with the ML criterion (as in [6]) allows dividing by two the gap of about 50 dB between the CR quantization performances obtained using the ML and the CR-MDL criteria. However, this last method leads to a very chaotic performance behaviour (this is due to the model parameter that changes with rate), and the remaining gap of 25 dB is still large.

4) Simulations: In our experiments we consider a so called *zero-rate distortion* D_0 defined as

$$\log_2 D_0 = \frac{2}{k}\left(\psi^{\text{flex}}(\theta, \mathbf{s}) + R_{\text{mod}}^{\text{fix}}\right), \tag{43}$$

where $R_{\text{mod}}^{\text{fix}}$ is the fixed rate used for transmission of the model (components and gains). Let $R_{\text{tot}} = R + R_{\text{mod}}^{\text{fix}}$ be the total rate. It is easy to see from Eq. (34) that the zero-rate distortion D_0 corresponds to (HR asymptotic) distortion for $R_{\text{tot}} = 0$. It is in fact a measure of asymptotic coding performance.

a) Shape factors and number of GGSMM components: We have computed zero-rate distortion for all three settings (i)-(iii), for shape factors $\nu = 1, 1.5, 2$, and for the number of components Q varying as $\log_2 Q = 0, 1, \ldots, 9$. The results are shown on the top row of figure 2. First, we see again that, as compared to the ML criterion, the CR-MDL criterion significantly improves and stabilizes the performance in the CR case. Second, the CR-ML performance closely approaches the CR-OPT performance for Laplacian distribution ($\nu = 1$) with many components, thus the mismatch between the ML and the CR-MDL criteria is lowest for this model. This result indicates that the mixture of Laplacian distributions with many components is probably the most appropriate model for speech

among all the models considered. Finally, while the Laplacian distribution leads to the best results (i.e., the best RD tradeoff) for the CE-OPT case for all values of Q, in the CR-OPT case Laplacian distribution gives the best results for small values of Q, but this tendency inverts for large values of Q. The heavy tails of the Laplacian distribution aid in the quantization of outliers for low Q, while the smooth shape around the mode of a Gaussian facilitates accurate modeling of an arbitrary smooth distribution at high Q. These results show that there are some practical situations, where using non-Gaussian models can be beneficial, as compared to Gaussian.

b) Coding: Here we check whether the effective quantization performances approach theoretically predicted ones at high-rates, and also we would like to see what happens at low rates. For that we perform with the GGSMM-based coding scheme the experiments similar to those reported on Fig. 1 for all settings (i)-(iii), number of components $Q = 64$, and for shape factors $\nu = 1, 1.5, 2$. The results are shown on the bottom row of figure 2. Note that, these results, in contrast to those from Fig. 1, are plotted in the absolute scale, and not relatively to the CE-OPT theoretical performance. In fact, we see that real quantization results approach their high rate asymptotics in all cases. Again, in line with what was observed in the synthetic data case (see Fig. 1), the asymptotic behavior is quite poor in the CR-ML case, notably for $\nu = 1.5$ or 2, and usage of the optimal criterion allows stabilizing it. Finally, while quantization results approach their asymptotics only for high rates, we see that for low rates (e.g., 5 - 10 bits per sample) the CR-MDL criterion outperforms systematically the ML criterion in the CR case.

5) Summary: In the CR case the CR-MDL criterion outperforms systematically the ML criterion (used in [6]) for high and moderately low rates (see the bottom row of figure 2). As compared to quantization using mixtures of Gaussian distributions with HR-optimally estimated parameters [7], [11], using mixtures of Laplacian distributions is beneficial for a small number of components (up to 16) in the CR case (see the top middle row of figure 2) and for any number of components we have tested in the CE case (see the top right row of figure 2). Thus, both the optimal estimation criteria and non-Gaussian modeling have their advantages for this task.

V. Discussion and Conclusion

We have proposed a framework of asymptotically optimal model estimation for quantization. This framework generalizes previous works to a wider family of model distributions, including non-Gaussian ones. We have evaluated the proposed estimation criteria and quantization schemes on synthetic data and speech MLT coefficients. Experiments showed that in the CR case the proposed CR-MDL criterion outperforms the ML criterion in all cases, thus compensating for the mismatch between model and data distributions.

It should be noted that such a "suboptimality" of the ML criterion for quantization in the CR case is related with other works and remarks in the literature. For example, Samuelsson [8] has tuned some factor (that equals to $\sqrt{3}$ according to theory) for his GMM-based quantization scheme[10] so that to optimize the performance. While no motivation was given in [8] for this tuning, our frameworks provides an obvious one. In fact, this factor scales with model standard deviations, and the goal of this tuning was to compensate for the mismatch between model and data distribution.

Our experiments on quantization of MLT speech coefficients with flexible quantizers based on such non-Gaussian models (e.g., scaled mixtures of Laplacian distributions) show that they can be advantageous, as compared to Gaussian models. The advantage of Laplacian distributions for speech was already shown for other applications (e.g., speech enhancement [26] and source separation [27]), and we confirm it for the coding application.

As for further research, an interesting direction would be to develop practical flexible model-based quantizers for hybrid rate constraints in-between CR and CE (e.g., as in [3]) and to derive corresponding optimal model estimation criteria. A practical advantage of such quantizers is that they would be able to avoid the most severe outliers in distortion of CR quantizers and the outliers in rate of the CE quantizers.

Appendix A
Derivation of the empirical mismatched operational RDF (Eqs (16), (17), (18))

Let $\mathbf{s} = \{s^n\}_{n=1}^N$ a sequence of vectors to quantize. Let $\mathcal{A} = \{A_m\}_{m\in\mathbb{Z}}$ be a partition of $\mathbb{R}^k$ into half-open cubes of side length $\varepsilon > 0$:

$$A_m = \left\{ s \in \mathbb{R}^k \,\middle|\, J_i(m) \le \frac{s_i}{\varepsilon} < J_i(m)+1, i = 1, \dots, k \right\},$$

where $J_i(m) = [J_1(m), \dots, J_k(m)]$ is a bijective mapping between $\mathbb{Z}$ and $\mathbb{Z}^k$. We consider a histogram-based empirical density estimate with pdf

$$\hat{p}_S(s|\mathbf{s},\varepsilon) = \frac{1}{\varepsilon^k N} \sum\nolimits_{n=1}^{N} \sum\nolimits_{m\in\mathbb{Z}} \mathbf{1}_{A_m}(s^n), \qquad (44)$$

where $\mathbf{1}_A(\cdot)$ is the indicator function of a subset $A \subset \mathbb{R}^k$.

For the results on mismatched operational RDFs (11), (12), (13) to be applicable to the data and model distributions with pdfs $\hat{p}_S(s|\mathbf{s},\varepsilon)$ and $f_S(s|\theta)$, one needs to assure the sufficient conditions of Theorem 2 of [17] and of Theorem 2 of [16] are satisfied.

Sufficient conditions of Theorem 2 of [17] are:

- CE.1 Differential entropy $h(f_S,\theta) = -\int_{\mathbb{R}^k} f_S(s|\theta)\log f_S(s|\theta)ds$ exists and it is finite.
- CE.2 For every optimal quantizer Q its entropy $H_{f_S,\theta}(Q) = -\sum_j \int_{V_j} f_S(s|\theta)ds \log \int_{V_j} f_S(s|\theta)ds$ (where V_j denotes quantization cells and the summation is over all cells) exists and it is finite.
- CE.3 $f_S(s|\theta) = 0$ implies $p_S(s) = 0$ for all s.
- CE.4 $p_S(s)/f_S(s|\theta)$ is bounded.

Sufficient conditions of Theorem 2 of [16] are:

- CR.1 There exist $\delta > 0$ such that $\int_{\mathbb{R}^k} \|s\|^{r+\delta}(p_S(s) + f_S(s|\theta))ds < +\infty$.

[10] From [8]: "The factor c_c in the encoding and decoding was experimentally tuned to maximize either SNR or PESQ for each model at rate 2 (the same factor was used at the other rates)."

CR.2 $p_S(s)$ and $f_S(s|\theta)$ satisfy

$$\frac{\int_{\mathbb{R}^k} p_S(s) f_S(s|\theta)^{-\frac{r}{k+r}} ds}{\left(\int_{\mathbb{R}^k} f_S(y|\theta)^{\frac{k}{k+r}} dy\right)^{-\frac{r}{k}}} < +\infty \qquad (45)$$

We assume that for every θ and all the CE quantizers Q we consider here conditions CE.1 and CE.2 are satisfied. We also assume the model pdf $f_S(s|\theta)$ to be continuous and positive in $\mathbb{R}^k$. Finally, we assume that there exist $\delta > 0$ such that $\int_{\mathbb{R}^k} \|s\|^{r+\delta} f_S(s|\theta) ds < +\infty$ and that $\int_{\mathbb{R}^k} f_S(y|\theta)^{\frac{k}{k+r}} dy < +\infty$. With these assumptions and because of the fact that $\hat{p}_S(s|\mathbf{s}, \varepsilon)$ has a bounded support in $\mathbb{R}^k$ conditions CE.3, CE.4, CR.1 and CR.2 are satisfied.

Thus, we can write the mismatched operational RDFs equations (11), (12), (13) for $\hat{p}_S(s|\mathbf{s}, \varepsilon)$ and $f_S(s|\theta)$. Doing so, and tending ε to zero we obtain, due to continuity of $f_S(s|\theta)$, the empirical mismatched operational RDFs expressed by (16), (17) and (18).

APPENDIX B
DERIVATION OF THE CR MISMATCHED OPERATIONAL RDF FOR FLEXIBLE QUANTIZERS (EQS (34), (36))

Since, because of orthogonality of U_n, the MSE distortion measure is invariant under transform $U_n^T(\cdot - \mu_n)$, we can consider quantization of transformed data vectors $\mathbf{y} = \{y^n\}_{n=1}^N$ ($y^n = U_n^T(s^n - \mu_n)$) instead of quantization of $\mathbf{s} = \{s^n\}_{n=1}^N$, without any loss of generality.

Let $R_{n,i}$ be the rate spent for quantization of i-th dimension of vector y^n, and $D_{n,i}$ be the corresponding expected distortion. Rewriting the CR empirical operational RDF defined by Eqs (16) and (18) in the particular case of $r = 2$, $k = 1$, $C_{r,k} = C_s$, $R = R_{n,i}$, $D = D_{n,i}$, $N = 1$, and $s^1 = y_i^n$, we have the following relation between $R_{n,i}$ and $D_{n,i}$:

$$R_{n,i} = -\frac{1}{2}\log_2 D_{n,i} + \frac{1}{2}\log_2 \frac{C_s f_{Y_i^n}(y_i^n)^{-\frac{2}{3}}}{\left(\int_{\mathbb{R}} f_{Y_i^n}(z_i)^{\frac{1}{3}} dz_i\right)^{-2}}, \quad (46)$$

where $f_{Y_i^n}(y_i)$ is given by (28).

The overall average distortion D (per dimension) can be expressed as

$$D = \frac{1}{kN}\sum_{i=1}^{k}\sum_{n=1}^{N} D_{n,i}. \qquad (47)$$

Using (47), $D_{n,i}$ expressed via (46), $R_{n,i}$ expressed via (33), and expression (28) for $f_{Y_i^n}(y_i)$, we obtain equation (34) with $\psi_{\mathrm{CR}}^{\mathrm{flex}}(\theta, \mathbf{s})$ defined by (36).

APPENDIX C
MISCELLANEOUS DETAILS ON IMPLEMENTATION

A. CR quantization with GGD

Here we consider the case of scalar CR quantization based on a GGD, i.e., when $\eta(\cdot) = f_{\mathrm{GGD}}(\cdot|\nu, 1)$ (see Sec. III-B). To implement the optimal scalar compressor $\xi_{\frac{1}{3}}(\cdot)$ and expander $\xi_{\frac{1}{3}}^{-1}(\cdot)$ (see Sec. III-C2) in this case, one only needs to compute the cdf of the corresponding GGD and its inverse. This simplification results from the fact that for the GGD $\eta_{\frac{1}{3}}(x) = 3^{-1/\nu}\eta(3^{-1/\nu}x)$ (note that this is not a general property).

The cdf of the centered GGD with unit variance and shape parameter ν can be written as:

$$\xi(x) = \frac{1}{2}\left[1 + \mathrm{sign}(x)\gamma\left(\frac{1}{\nu}, (\alpha(\nu)|x|)^\nu\right)\right], \qquad (48)$$

where $\gamma(a, y)$ is the *lower incomplete Gamma function* defined as (we are using the Matlab definition of this function):

$$\gamma(a, y) = \frac{1}{\Gamma(a)}\int_0^y t^{a-1}e^{-t}\, dt. \qquad (49)$$

The inverse cdf is computed similarly using the *inverse upper incomplete Gamma function* $\gamma^{-1}(a, y)$, i.e., the inverse of $\gamma(a, y)$ w.r.t. y. In our Matlab implementation we used `gammainc` and `gammaincinv` functions to compute $\gamma(a, y)$ and $\gamma^{-1}(a, y)$.

B. CR-MDL criterion optimization for GGD

In the case of the GGD model considered in the experimental section IV-A the term $\phi(\theta, \mathbf{s})$ (38) becomes:

$$\phi(\sigma, \mathbf{s}) = \chi(\nu)\sigma^2 \sum_{n=1}^{N} \exp\left[\frac{2}{3}\frac{\alpha(\nu)^\nu}{\sigma^\nu}|s^n|^\nu\right], \qquad (50)$$

where $\chi(\nu) = 3^{\frac{2}{\nu}}\left(\frac{2\Gamma(1/\nu)}{\nu\alpha(\nu)}\right)^2$ is a constant that is independent on σ. To minimize this term we use either Newton's method or a gradient descent algorithm w.r.t. $\log\sigma$, instead of σ, since that incorporates a non-negativity constraint in the optimization. The first and the second derivatives of $\phi(\sigma, \mathbf{s})$ w.r.t. $\log\sigma$ needed for this optimization can be expressed as:

$$\frac{\partial}{\partial \log\sigma}\phi(\theta, \mathbf{s}) = \chi(\nu)\sigma^2\left[2\zeta_0 - \nu\,\zeta_1\right], \qquad (51)$$

$$\frac{\partial}{\partial^2 \log\sigma}\phi(\theta, \mathbf{s}) = \chi(\nu)\sigma^2\left[2\zeta_0 + \nu(\nu-3)\zeta_1 + \nu^2\zeta_2\right], \quad (52)$$

where

$$\zeta_l = \sum_{n=1}^{N}\left[\frac{2}{3}\frac{\alpha(\nu)^\nu}{\sigma^\nu}|s^n|^\nu\right]^l \exp\left[\frac{2}{3}\frac{\alpha(\nu)^\nu}{\sigma^\nu}|s^n|^\nu\right], \quad l = 0, 1, 2.$$

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable comments.

REFERENCES

[1] R. M. Gray, *Source coding theory*. Kluwer Academic Press, 1990.
[2] P. L. Zador, "Asymptotic quantization error of continuous signals and the quantization dimension," *IEEE Transactions on Information Theory*, vol. IT-28, no. 2, pp. 139–148, Mar. 1982.
[3] R. M. Gray, T. Linder, and J. T. Gill, "Lagrangian vector quantization with combined entropy and codebook size constraints," *IEEE Transactions on Information Theory*, vol. 54, no. 5, pp. 2209–2242, May 2008.
[4] P. Frossard, P. Vandergheynst, R. M. Figueras I Ventura, and M. Kunt, "A posteriori quantization of progressive matching pursuit streams," *IEEE Transactions on Signal Processing*, vol. 52, no. 2, pp. 525–535, 2004.
[5] E. Ravelli and L. Daudet, "Embedded polar quantization," *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 657–660, Oct.s 2007.

[6] A. Subramaniam and B. Rao, "PDF optimized parametric vector quantization of speech line spectral frequencies," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 2, pp. 130–142, March 2003.
[7] D. Zhao, J. Samuelsson, and M. Nilsson, "On entropy-constrained vector quantization using Gaussian mixture models," *IEEE Transactions on Communications*, vol. 56, no. 12, pp. 2094–2104, 2008.
[8] J. Samuelsson, "Waveform quantization of speech using Gaussian mixture models," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04)*, vol. 1, May 2004, pp. 165–168.
[9] M. Li and W. B. Kleijn, "A low-delay audio coder with constrained-entropy quantization," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'07)*, Nov. 2007, pp. 191–194.
[10] A. Ozerov and W. B. Kleijn, "Flexible quantization of audio and speech based on the autoregressive model," in *IEEE Asilomar Conference on Signals, Systems, and Computers (Asilomar CSSC'07)*, Nov. 2007, pp. 535–539.
[11] E. R. Duni and B. D. Rao, "A high-rate optimal transform coder with Gaussian mixture companders," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 770–783, Mar 2007.
[12] A. Ozerov and W. B. Kleijn, "Optimal parameter estimation for model-based quantization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'09)*, April 2009, pp. 2497–2500.
[13] W. B. Kleijn and A. Ozerov, "Rate distribution between model and signal," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'07)*, 2007, pp. 243–246.
[14] P. Grünwald, "Model selection based on minimum description length," *Journal of Mathematical Psychology*, vol. 44, pp. 133–152, 2000.
[15] R. M. Gray and T. Linder, "Results and conjectures on high rate quantization," in *Proceedings of the 2004 IEEE Data Compression Conference*, 2004, pp. 3–12.
[16] J. Bucklew, "Two results on the asymptotic performance of quantizers," *IEEE Transactions on Information Theory*, vol. 30, no. 2, pp. 341–348, 1984.
[17] R. M. Gray and T. Linder, "Mismatch in high-rate entropy-constrained vector quantization," *IEEE Transactions on Information Theory*, vol. 49, no. 5, pp. 1204–1217, 2003.
[18] J. Rissanen, "Modeling by the shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.
[19] A. Barron and T. M. Cover, "Minimum complexity density estimation," *IEEE Transactions on Information Theory*, vol. 37, no. 4, pp. 1034–1054, 1991.
[20] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Kluwer Academic Press, 1992.
[21] A. Gersho, "Asymptotically optimal block quantization," *IEEE Transactions on Information Theory*, vol. 25, pp. 373–380, 1979.
[22] A. D. Subramaniam, W. R. Gardner, and B. D. Rao, "Low-complexity source coding using Gaussian mixture models, lattice vector quantization, and recursive coding with application to speech spectrum quantization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 524–532, 2006.
[23] T. D. Lookabaugh and R. M. Gray, "High-resolution quantization theory and the vector quantizer advantage," *IEEE Transactions on Information Theory*, vol. 35, no. 5, pp. 1020–1033, 1989.
[24] T. Linder, R. Zamir, and K. Zeger, "High-resolution source coding for non-difference distortion measures: Multidimensional companding," *IEEE Transactions on Information Theory*, vol. 45, no. 2, pp. 548 – 561, March 1999.
[25] M. Oger, S. Ragot, and M. Antonini, "Low-complexity wideband LSF quantization by predictive KLT coding and generalized Gaussian modeling," in *EUSIPCO, 14th European Signal Processing Conference*, Florence, Italia, Sept. 2006.
[26] R. Martin and C. Breithaupt, "Speech enhancement in the DFT domain using Laplacian speech priors," in *Proc. 8th Intern. Workshop on Acoustic Echo and Noise Control (IWAENC)*, vol. 8-11, Kyoto, Japan, 2003, pp. 87–90.
[27] E. Vincent, "Complex nonconvex lp norm minimization for underdetermined source separation," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA'07)*, 2007, pp. 430–437.
[28] B.-H. Juang and L. Rabiner, "The segmental K-means algorithm for estimating parameters of hidden Markov models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 38, pp. 1639–1641, 1990.
[29] A. P. Dempster, N. M. Laird, and D. B. Rubin., "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, pp. 1–38, 1977.

Alexey Ozerov holds a Ph.D. in Signal Processing from the University of Rennes 1 (France). He worked towards this degree from 2003 to 2006 in the labs of France Telecom R&D and in collaboration with the IRISA institute. Earlier, he received an M.Sc. degree in Mathematics from the Saint-Petersburg State University (Russia) in 1999 and an M.Sc. degree in Applied Mathematics from the University of Bordeaux 1 (France) in 2003. From 1999 to 2002, Alexey worked at Terayon Communicational Systems (USA) as a R&D software engineer, first in Saint-Petersburg and then in Prague (Czech Republic). He was for one year (2007) in Sound and Image Processing Lab at KTH (Royal Institute of Technology), Stockholm, Sweden, and for one year and half (2008-2009) in Télécom ParisTech / CNRS LTCI - Signal and Image Processing (TSI) Department. Now he is with METISS team of IRISA / INRIA - Rennes as a Post-Doc researcher.

Bastiaan Kleijn is Professor of Electronic Engineering at Victoria University of Wellington, New Zealand since 2010. He is also a Professor at the School of Electrical Engineering at KTH (the Royal Institute of Technology) in Stockholm, Sweden, which he joined in 1996 and where he was until recently Head of the Sound and Image Processing Laboratory. He holds a Ph.D. in Electrical Engineering from Delft University of Technology (Netherlands), a Ph.D. in Soil Science and an M.S. in Physics, both from the University of California, Riverside, and an M.S. in Electrical Engineering from Stanford University. He worked on speech processing at AT&T Bell Laboratories from 1984 to 1996. He was a founder of Global IP Solutions, which was acquired by Google in 2010. He is on the Editorial Board of Signal Processing and has been on the Boards of IEEE Transactions of Speech and Audio Processing, IEEE Signal Processing Letters, IEEE Signal Processing Magazine, and the EURASIP Journal of Applied Signal Processing. He has been a member of several IEEE technical committees, and a Technical Chair of EUSIPCO 2010, ICASSP-99, the 1997 and 1999 IEEE Speech Coding Workshops, and a General Chair of the 1999 IEEE Signal Processing for Multimedia Workshop. He is a Fellow of the IEEE.